Eigenvalue routines for overlap fermions

June 20

Nigel Cundy
Weonjong Lee, Seonghee Kim

Introduction

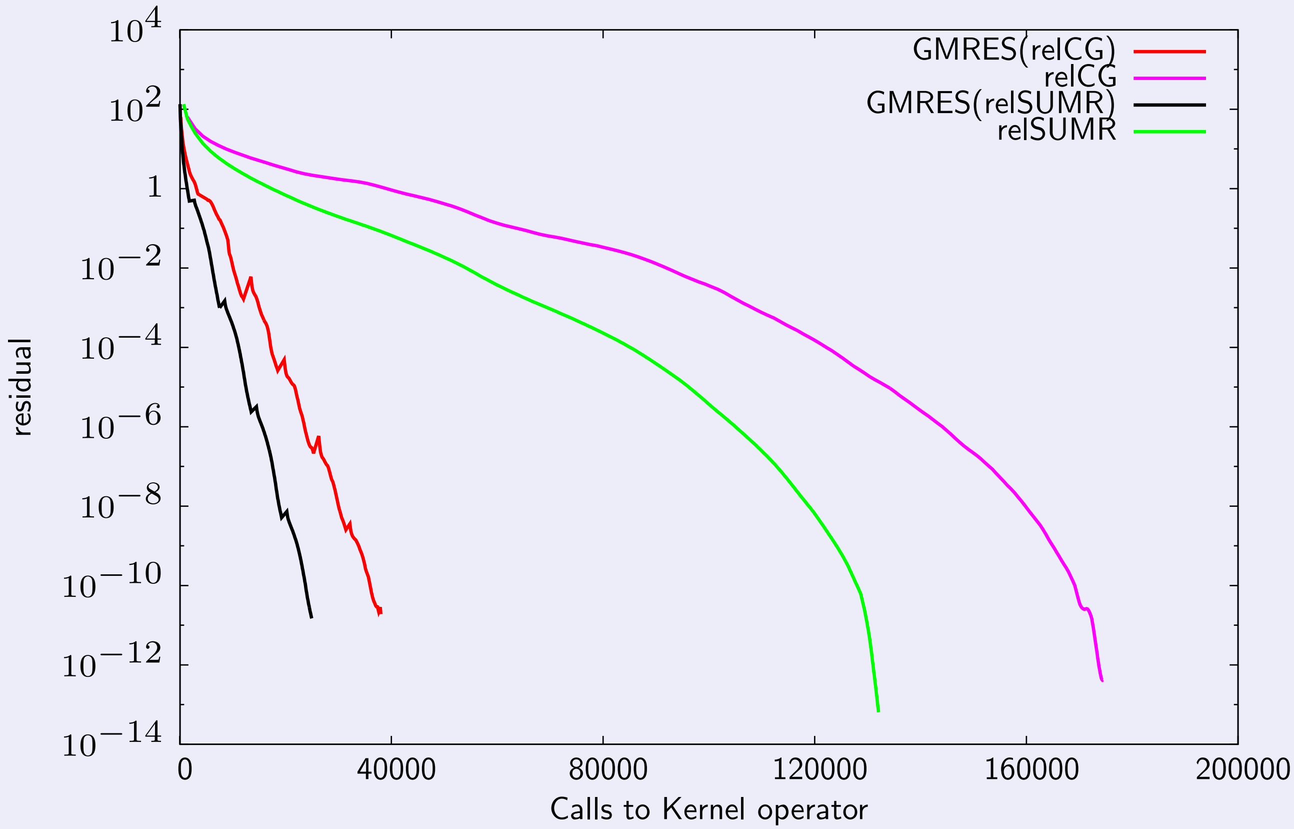
- The overlap operator is (theoretically) the cleanest Dirac operator available in lattice QCD – everybody should be using it
- It is also the most expensive Dirac operator available, and the most difficult algorithmically, and it is unlikely that the advantages of exact chiral symmetry in the massless limit outweigh the costs – nobody should be using it
- Nonetheless, it is important to confirm our calculations using different methods
- Some studies, for which chiral perturbation theory cannot compensate for the symmetry breaking (QCD Vacuum? Chiral Magnetic Effect?) may be easier or more accurate with overlap fermions
- Reducing the cost of overlap simulations is thus a worthy area of study

$$D = \frac{1}{2}(1 + \mu + (1-\mu)\gamma_5 \textsf{sign}(K))$$

- K is some Hermitian Kernel operator – say $\gamma_5(D_W - 1)$.
- μ is a mass parameter
- Lots of theoretical advantages, mostly associated with an exact chiral symmetry as $\mu \to 0$.

- Five approaches to simulate the matrix sign function
  - Spectral Decomposition: $\text{sign}(K) = \sum_i |\psi_i\rangle\langle\psi_i|\text{sign}(\lambda_i)$.
  - Lanczos approach (I won't discuss further)
  - Polynomial Approximation (e.g. Chebychev)
  - Rational Approximation (e.g. Zolotarev)
  - Five Dimensional representation (I won't discuss further)
- The full spectral decomposition is impractical, but partial deflation is essential
- Rational approximations generally require fewer calls to K
- Polynomial approximations require less additional spinor algebra per call to K
- In most of these methods, it is much cheaper (perhaps a factor of 10) to calculate a low accuracy approximation to the matrix sign function compared to a high accuracy approximation
- Low accuracy sign functions only require single precision

- The goal when designing a routine for overlap fermions is to use as low accuracy approximation to the sign function as much as possible
- It is known how to do this for inversions:
  - Start with a high accuracy overlap operator, and gradually relax the accuracy until the last few calls are low accuracy
  - Use a low accuracy inversion as a preconditioner for a high accuracy inversion
- In total, we get at least a factor of 5 or 6 over the naive inversion.



- SUMR = Shifted Unitary Minimal Residual
  (the optimal Krylov subspace algorithm for overlap fermions).

- But what about the eigenvalues?
- Eigenvalues/vectors are needed in lattice QCD observables:
  - To deflate the inversion (low accuracy)
  - To reduce the measurement error of certain observable on each configuration (Low Mode Averaging, Truncated Eigenvalue Approximation) (high accuracy)
  - To directly calculate observables (e.g. Chiral Condensate, QCD vacuum) (high accuracy)

- The overlap operator is shifted unitary – a normal operator
- The eigenvalues lie on a circle in the complex plane
- Real eigenvalues ψ_0, ψ_1 at $\lambda = \pm 1$
- Other eigenvalues in complex conjugate pairs $\lambda_\pm = \lambda^2 \pm i\lambda\sqrt{1-\lambda^2}$
- The Hermitian overlap operator $\gamma_5 D$ has eigenvalues $\pm\lambda$ with eigenvectors $\psi_\pm$
- The squared Hermitian overlap operator $D^\dagger D$ has degenerate non-zero eigenvalues
- $\gamma_5\psi_{\pm,i}$ is a linear combination of $\psi_{+,i}$ and $\psi_{-,i}$
- we can construct the eigenvectors from just about any non-trivial function of γ_5 and sign(K)
- The eigenvectors of D, $\gamma_5 D$, $D^\dagger D$ are independent of the quark mass

- Deflation constructs a preconditioner or a starting guess for the inversion using the smallest eigenvalues and eigenvectors
- The condition number of the operator improves by the ratio of the smallest and largest eigenvalues you calculate
- In typical lattice simulations, possible to get a factor of > 5 gain
- This is perhaps slightly old technology (Multigrid?) but still useful in some circumstances
- Obviously larger lattice, mixed action approaches require more eigenvalues so the problem becomes harder

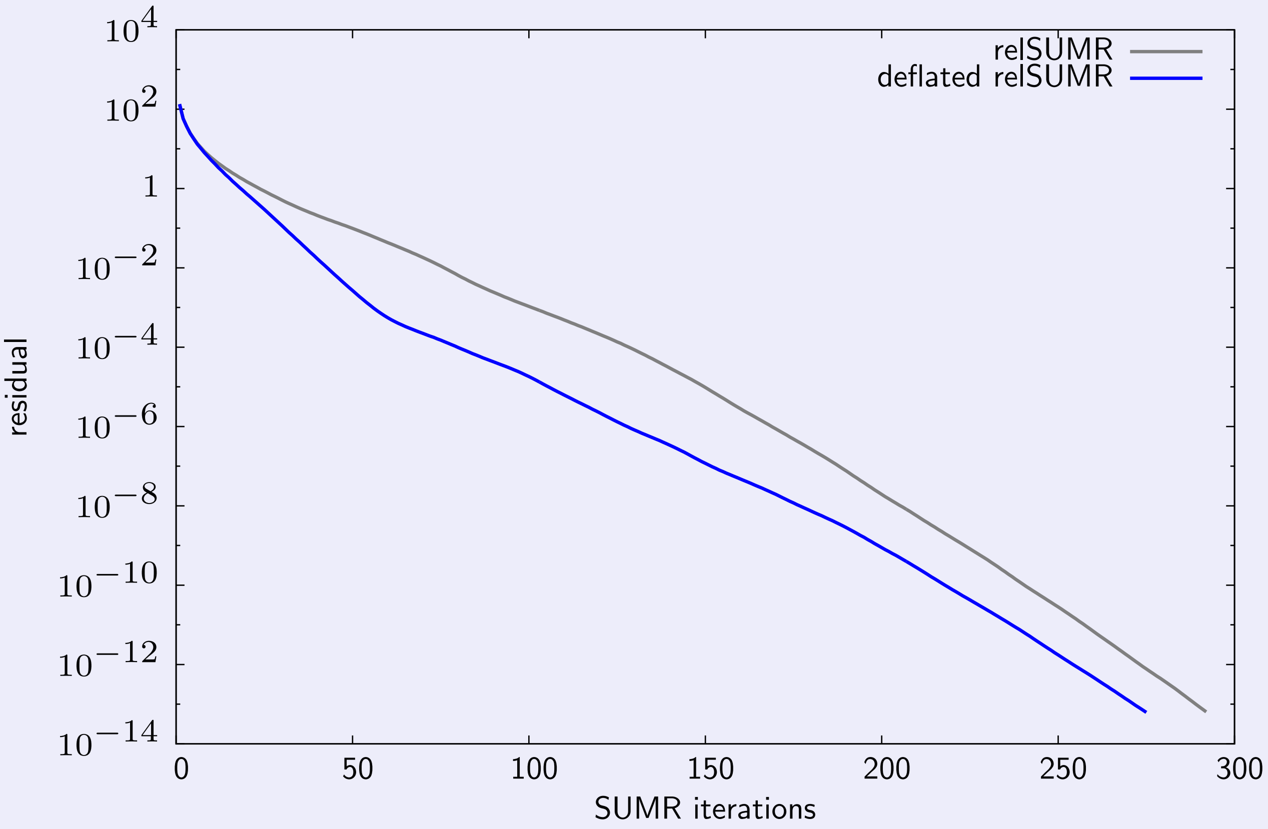
- Method 1: $Ax = b$, for any A and routine
  - We construct an initial guess x_0 from the eigenvectors of A

$$x_0 = \sum_i \alpha_i \psi_i$$

  - We choose α_i to minimise the residual $\|Ax_0 - b\|$.
  - Construct the ‘eigenvector’ basis so $\psi_i A^\dagger A \psi_j$ is diagonal

$$\alpha_i = \frac{(\psi_i, Ab)}{(\psi_i, A^\dagger A \psi_i)}$$

  - This method accelerates the inversion until the inversion residual $\sim$ the eigenvector residual.
  - After that, the inversion proceeds at the undeflated rate.
  - Deflate the low accuracy preconditioner – only need low accuracy eigenvalues.


relSUMR
deflated relSUMR
residual
10^{4}
10^{2}
1
10^{-2}
10^{-4}
10^{-6}
10^{-8}
10^{-10}
10^{-12}
10^{-14}
0
50
100
150
200
250
300
SUMR iterations

- Method 2: Invert $Ax = b$, A Hermitian and positive definite

$$x = P\frac{1}{PAP}P$$
$$P = (1 - \sum_i \psi_i \psi_i^\dagger) + \sum_i \psi_i \psi_i^\dagger \frac{\sqrt{c}}{\sqrt{\lambda_i}}$$

  - The eigenvalues/eigenvector (λ_i/ψ_i) need only be calculated to a very low accuracy to achieve the full gain
  - If the cost of applying the pre-conditioner $\approx$ the cost of applying A, this method may not be useful
  - For overlap fermions, who cares?
  - But only useful for the CG inversion of $1/D^\dagger D$
  - SUMR, multishift etc., cannot use this preconditioning
- Method 2 tends to be better, where it can be used.

- Here we want to consider four eigenvalue routines for overlap fermions
  - eigSUMR
  - Explicitly restarted Unitary Lanczos
  - Jacobi-Davidson
  - Zolotarev

- The SUMR (shifted Unitary Minimal Residual) routine is a way of constructing an Arnoldi Basis for unitary and shifted unitary operators using short recurrences
- Hermitian/Lanzos $\Leftrightarrow$ Shifted Unitary/SUMR
- It generates a series of orthonormal vectors q_i, $i = 0, \dots m-1$ in the Krylov subspace $K_m(b) = \{b, Ab, A^2b, \dots, A^{m-1}b\}$
- $U = \gamma_5 \text{sign}(K)$ is unitary – applicable for $\frac{2D}{1-\mu} = \frac{1+\mu}{1-\mu} + U$

```
$q_0 = \tilde{q}_0 = b/\|b\|$
for $j$ in $0, 1, 2, 3, 4, \dots$; do
    $u = Uq_j$
    $\gamma_j = -(\tilde{q}_j, u); \quad \sigma_j = \sqrt{1 - |\gamma_j|^2}$
    $q_{j+1} = \frac{1}{\sigma_j}(u + \gamma_j \tilde{q}_j); \quad \tilde{q}_{j+1} = \sigma_j \tilde{q}_j + \gamma_j^* q_{j+1}$
done
```

- This recurrance can be used for a minimal residual inversion routine
- For a inversion, what accuracy η do we need to calculate U at each iteration j to maintain a desired final accuracy for the inversion $\|r_j\| \equiv \|Ax_j - b\| \leq \epsilon_A \|b\|$?

$$||b - Ax_k|| \leq ||r_k - (b - Ax_k)|| + ||r_k||$$

- We want to control the residual gap, $||r_k - (b - Ax_k)||$, so that it is smaller than the target accuracy
- The optimal result for a minimal residual agorithm is

$$\eta_j = \epsilon_A \|b\| / \|r_j\|$$

- Now suppose we want to use this subspace to calculate n eigenvalues where we can at most store m vectors?

```
$q_0 = \tilde{q}_0 = b/\|b\|; \quad \mathbf{v} = 0; \quad \mathbf{v}^D = 0; \quad k = 0$
for $j$ in $0, 1, 2, 3, 4, \ldots;$ do
    $u = Uq_j; \quad v_k = q_j; \quad v_k^D = \frac{(1-\mu)}{2}u + \frac{(1+\mu)}{2}q_j$
    $k = k + 1$
    if $(k == m)$; then
        Diagonalise $M_{ij} = (v_i^D, v_j^D) + \delta_E(v_i, \gamma_5 v^D)$
        $k = n$
    end if
    $\gamma_j = -(\tilde{q}_j, u); \quad \sigma_j = \sqrt{1 - |\gamma_j|^2};$
    $q_{j+1} = \frac{1}{\sigma_j}(u + \gamma_j \tilde{q}_j); \quad \tilde{q}_{j+1} = \sigma_j \tilde{q}_j + \gamma_j^* q_{j+1}$
done
Diagonalise $M_{ij} = (v_i^D, v_j^D) + \delta(v_i, \gamma_5 v^D)$
```

- To remove degeneracies for the non-zero eigenvalues, we diagonalise

$$\begin{aligned} M_{ij} =& (v_i^D, v_j^D) + \delta_E(v_i, \gamma_5 v^D) \\ =& (v_i, (\gamma_5 D \gamma_5 D + \delta_E \gamma_5 D) v_j) \end{aligned}$$

  ($\delta_E =$ some small number).
- Obviously we can combine this with an inversion routine
- The basic idea is exactly the same as the eigCG algorithm
- We have called this routine eigSUMR
- Once the eigenvalues are good enough, we can start deflating
- Or we can run it as a stand alone eigenvalue solver – Unitary Lanczos

- The diagonalisation routine proceeds in two steps
  - We use an LDU decomposition to orthonormalise v; $(v_i, v_j) = (U^\dagger D U)_{ij}$ (U upper triangular, D diagonal $(1 or - 1)$
  - We use a spectral decomposition to diagonalise $(U^\dagger M U) = V^\dagger D' V$ (V unitary, D' diagonal)
  - The improved estimate of the eigenvectors is $v \to (V U^T)_{ji} v$
- It is useful to separately rediagonalise each non-zero eigenvector pair with respect to $(v_i, \gamma_5 v_j^D)$

- In principle, we do not need any additional calls to the overlap operator beyond the generation of the Krylov subspace to calculate the eigenvalues/eigenvectors
- In practice, the story is somewhat different
- If $v_i^D \approx Dv_i$ becomes too inaccurate, then the whole eigenvalue calculation disintegrates
- So how accurate do we need v^D for the eigenvalue calculation to remain stable?

- We use an approximate matrix sign function $\tilde{s}$ (with s the exact sign function)
- This leads to an approximate Dirac operator $\tilde{D}$ and an approximate $\tilde{v}^D$. We can write,

$$\tilde{v}^D = v^D + \delta,$$

- Our goal is to keep $\|\delta\|$ sufficiently small so that it has no significant effect on the estimate of the eigenvalue or the residual r^v

$$r_i^v = (v_i, (\gamma_5 D)^2 v_i) - (v_i, \gamma_5 D v_i)^2$$

- In inexact arithmetic

$$r^v = r^v_{\mathsf{true}} + \gamma_5\delta - v(v,\gamma_5\delta),$$

$$\|r^v\|^2 = \|r^v_{\mathsf{true}}\|^2 + (r^v_{\mathsf{true}}, (1 - vv^\dagger)\gamma_5\delta) + \\ ((1-vv^\dagger)\gamma_5\delta, r^v_{\mathsf{true}}) + (\delta, (1-vv^\dagger)\delta).$$

- The residual gap, $g = \|r^v\|^2 - \|r^v_{\mathsf{true}}\|^2$
- We want to keep $g < \epsilon^2$, where ϵ is the desired accuracy for the eigenvector.

$$g \leq 2\|r^v_{\mathsf{true}}\|\|\delta\| + \|\delta\|^2 < \epsilon^2.$$

This bound gives

$$\|\delta\| < \sqrt{\epsilon^2 + \|r^v_{\mathsf{true}}\|^2} - \|r^v_{\mathsf{true}}\|.$$

During each update of the eigenvectors, we know that

$$\begin{aligned} v_i \to & (VU^T)_{ij} v_j \\ \tilde{v}_i^D \to & (VU^T)_{ij} \tilde{v}_j^D = (VU^T)_{ij} v_j^D + (VU^T)_{ij} \delta_j^D, \end{aligned}$$

where δ_j^D is either

– The previously calculated error on an eigenvector

– Due to the application of D in the SUMR routine

- The new δ_i satisfies the bound

$$\|\delta_i\| < \sum_{j=1}^{m} |(UV)_{ij}| \|\delta_j^D\|.$$

- The rigorous bound computes the matrix sign function to an accuracy

$$\|\delta_j^D\| < \frac{1}{(m-n)k}\frac{\sqrt{\epsilon^2 + \|r^v_{\text{true},0}\|^2} - \|r^v_{\text{true},0}\|}{\max_{n<j\leq m, i<n'} |(UV)_{ij}|},$$

- We need to recalculate v^D to a high accuracy every k iterations
- $\|r^v_{\text{true},0}\|$ is the residual of the best converged eigenvector
- $\max_{ij}|(UV)_{ij}|$ may be estimated from the previous diagonalisations

- In practice, this is more conservative than we require, and we instead found the ‘sloppy bound’ works well

$$\|\delta_j^D\| < \frac{\chi}{\sqrt{(m-n)k}} \frac{\sqrt{\epsilon^2 + \|r^v_{\text{true},0}\|^2} - \|r^v_{\text{true},0}\|}{\max_{n'<j\leq m, i<n} |(UV)_{ij}|}$$

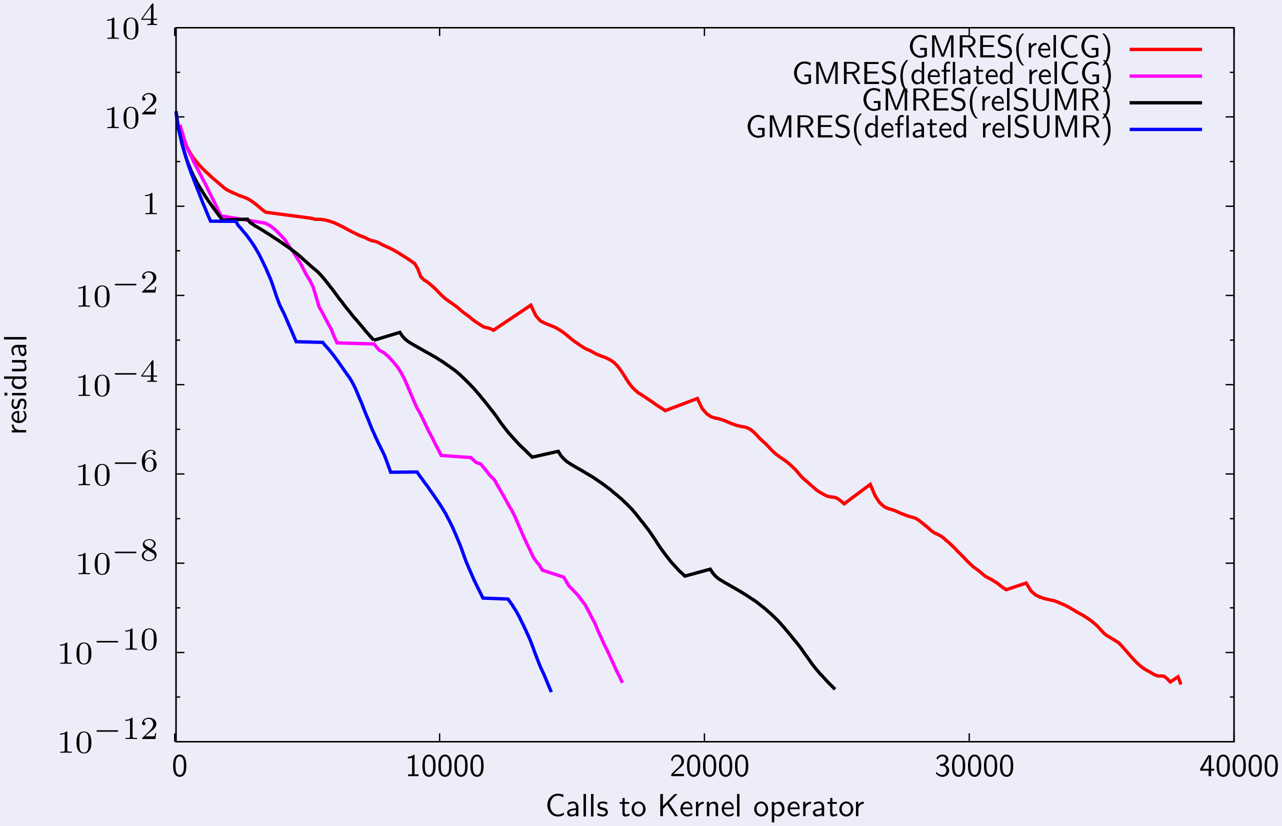
$$\chi = \begin{cases} \frac{1}{\sqrt{(m-n)k}} & \text{First few diagonalisations} \\ \sqrt{n} & \text{Subsequent diagonalisations} \end{cases} .$$

- This bound is $O(\epsilon)$
- We cannot usefully employ relaxation in eigSUMR to calculate eigenvectors to a high accuracy
- We can, of course, decrease the bound after each restart of the inverter – but this doesn't help so much in practice
- The SUMR q vectors quickly lose othogonolity when the overlap operator is calculated to a low accuracy
- If the SUMR vectors are not orthogonal, we need to project out the eigenvectors

$$v_j \to v_j - v_i(v_i, v_j)$$
$$v_j^D \to v_j^D - v_i^D(v_i, v_j)$$

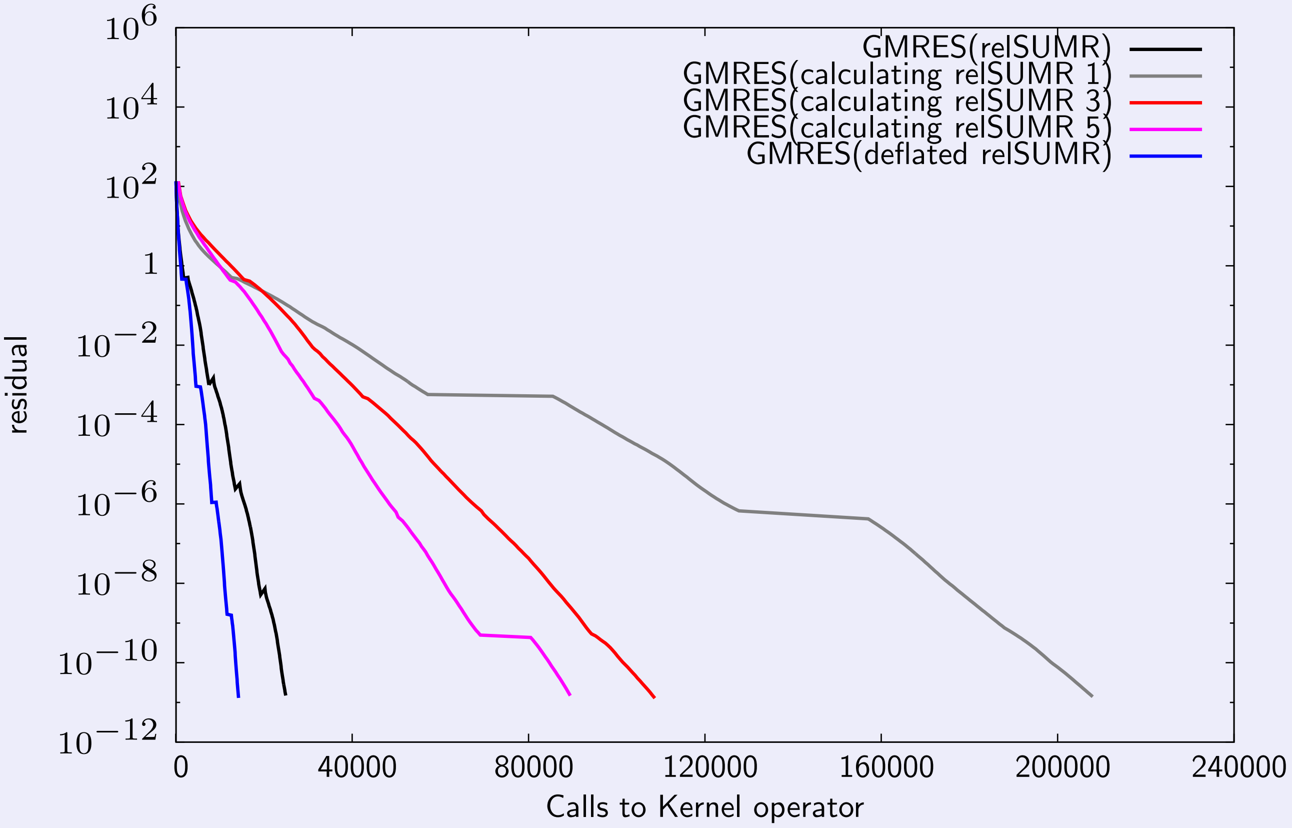
- We can quickly lose accuracy on v^D if there are near duplicate eigenvectors
- Need to continually check the accuracy of v^D and be prepared to recalculate it

- The gain from deflation:

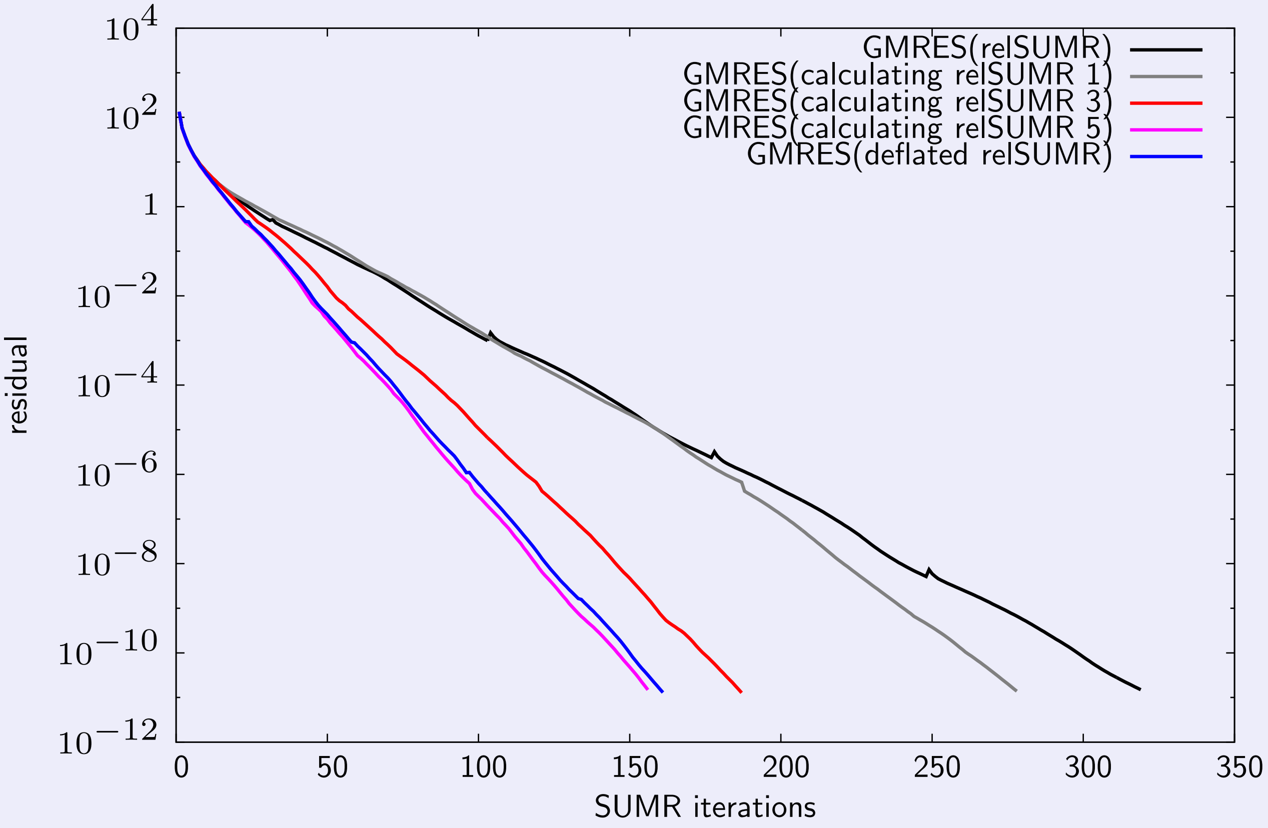


- All plots on an $8^3 \times 32$ dynamical overlap ensemble, lattice spacing ~ 0.12 fm, quark mass $\mu = 0.03$, $m_\pi \sim 460$ MeV

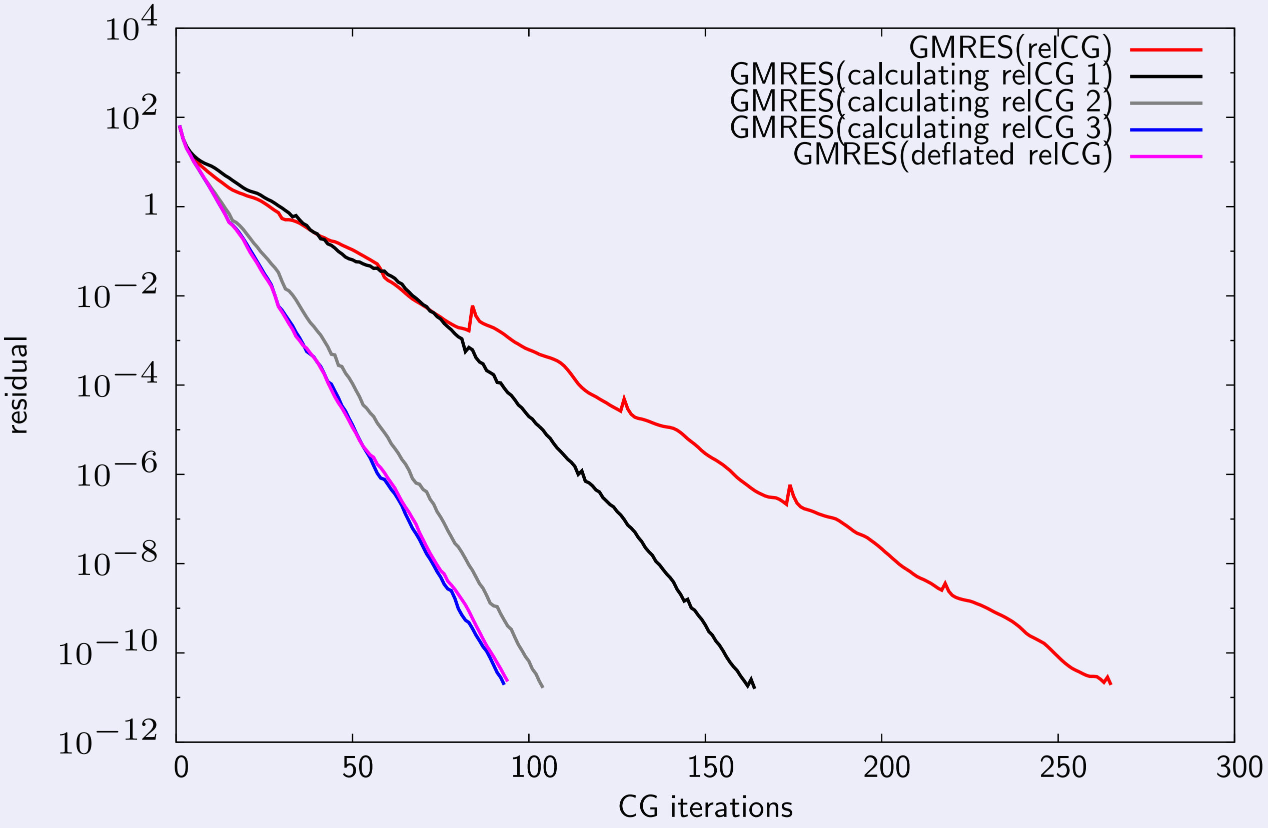
- The number of Wilson calls needed to calculate the eigenvectors:



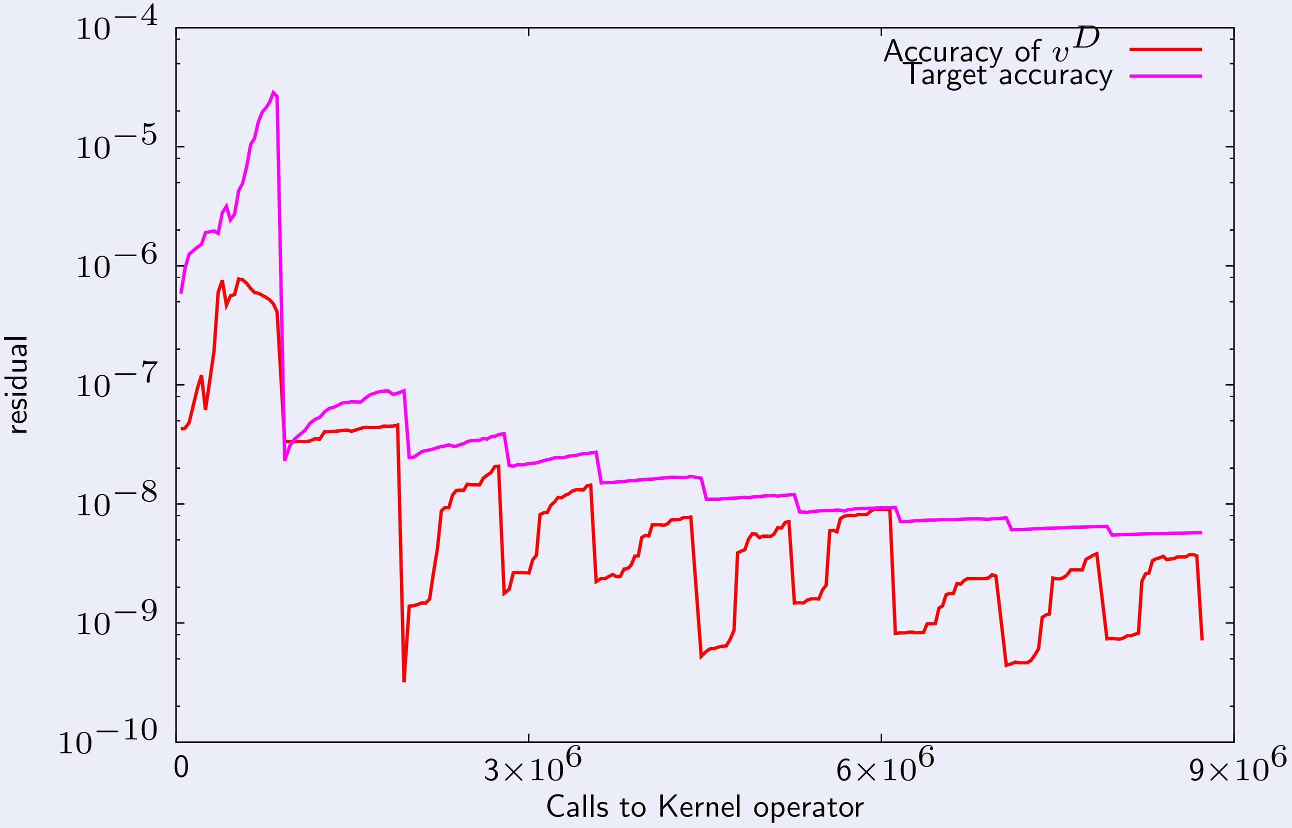
- The number of SUMR iterations needed to calculate the eigenvectors:



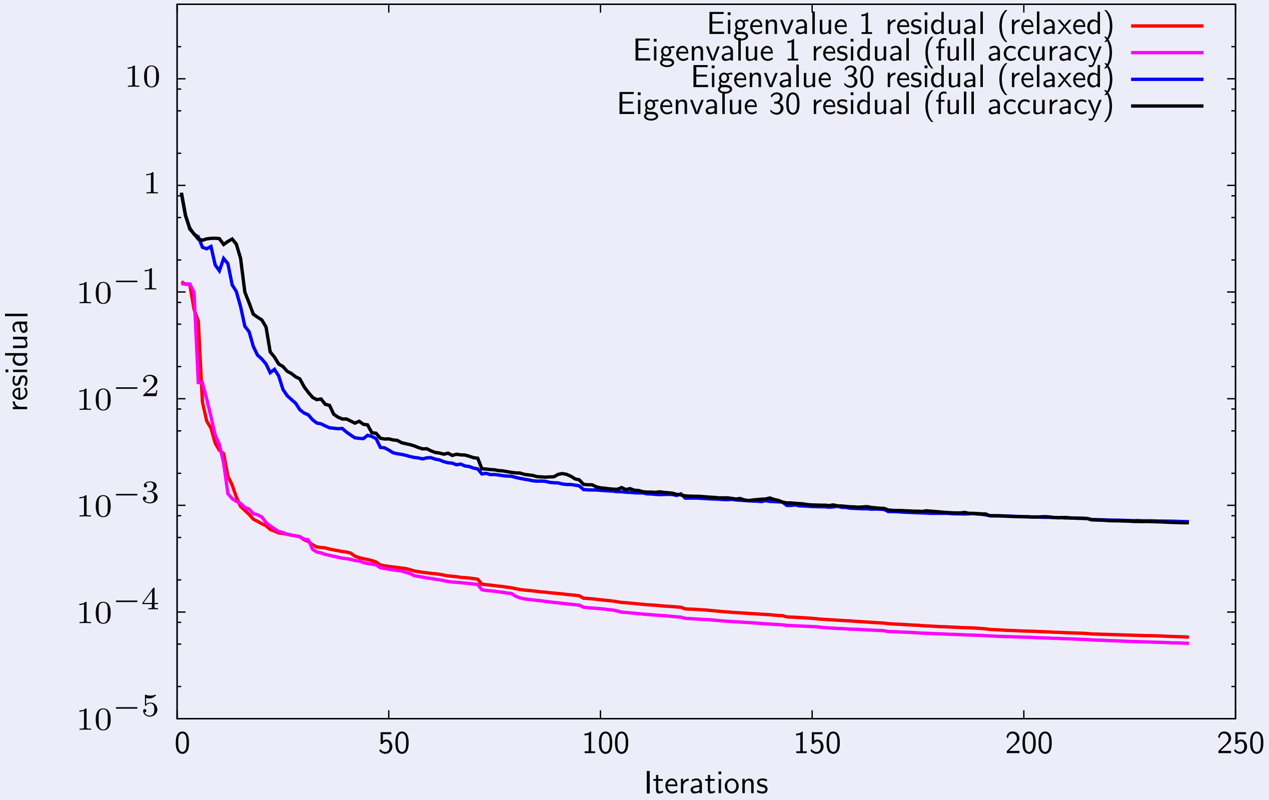
- The number of CG iterations needed to calculate the eigenvectors:



- The accuracy of the 29th V^D:

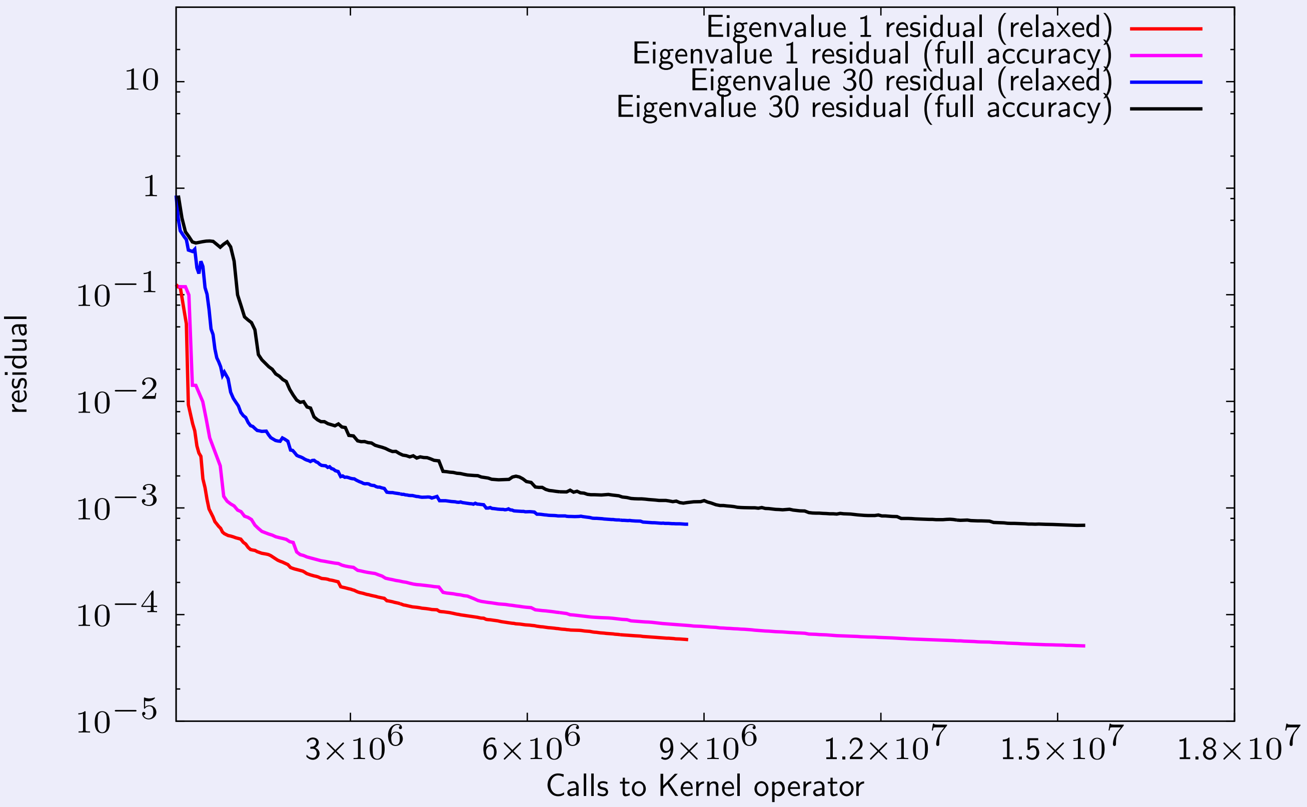


- The residuals compared to number of Arnoldi iterations



- Note that the convergence of the eigenvectors slows down dramatically after a certain residual
- EigSUMR/Unitary Lanczos are good for low accuracy eigenvectors; bad for high accuracy eigenvectors.

- The residual for the relaxed and unrelaxed unitary Lanczos routines compared to number of Wilson calls



Jacobi Davidson

- Suppose we have a guess of the lowest eigenvalue of a matrix A, $\mathbf{u}$.
- We define the Ritz estimate of the eigenvalue, λ and the residual, $\mathbf{r}$ as

$$\lambda = \frac{(\mathbf{u}, A\mathbf{u})}{(\mathbf{u}, \mathbf{u})} \qquad \mathbf{r} = A\mathbf{u} - \lambda\mathbf{u}.$$

- The true eigenvalue λ_* and eigenvector $\mathbf{u}_*$ satisfy the eigenvalue equation,

$$A\mathbf{u}_* = \lambda_*\mathbf{u}_*.$$

- We write

$$\mathbf{u}_* = \mathbf{u} + \mathbf{s},$$

- $\mathbf{s}$ is a small correction orthogonal to $\mathbf{u}$.

  $$(A-\lambda')(\mathbf{u}+\mathbf{s}) = (\lambda_* - \lambda')(\mathbf{u}+\mathbf{s}),$$

  or

  $$(A-\lambda')\mathbf{s} = -\mathbf{r} + (\lambda_* - \lambda')\mathbf{u} + (\lambda_* - \lambda')\mathbf{s},$$

  where λ' is any real number.
- We set λ' to the best estimate available for λ_*
- Neglect terms of O(s^2).
- Projecting into the subspace orthogonal to $\mathbf{u}$.

  $$(1-\mathbf{u}\mathbf{u}^\dagger)(A-\lambda')(1-\mathbf{u}\mathbf{u}^\dagger)\mathbf{s} = -\mathbf{r} + O(\mathbf{s}^2).$$

- This gives us our approximation to $\mathbf{s}$,

  $$\mathbf{s} \sim -\frac{1}{(1-\mathbf{u}\mathbf{u}^\dagger)(A-\lambda)(1-\mathbf{u}\mathbf{u}^\dagger)}\mathbf{r}.$$

- We now construct an orthonormal basis of vectors $V = \{\mathbf{v_1}, \mathbf{v_2}, \mathbf{v_3}, \ldots\}$, and find $V^A = \{\mathbf{v_1^A}, \mathbf{v_2^A}, \mathbf{v_3^A}, \ldots\} = AV$.
- We can obtain an improved estimate of the eigenvectors by diagonalising $E_{ij} = (v_i^A, v_j)$.
- By setting $\mathbf{v}_1 = \mathbf{u}$ and then $\mathbf{v}_2 = \mathbf{s}$, we can obtain the best estimate of the eigenvector in the subspace of $\mathbf{u}$ and $\mathbf{s}$.
- We repeat this process, until we have an accurate enough estimate of the eigenvector
- As long as we start close enough to the eigenvalue, it converges rapidly
- To expand this method for multiple eigenvectors, we use a subspace orthogonal to the eigenvectors already calculated

- This method puts the bulk of the work into inversions
- For overlap fermions, we no how to do an inversion efficiently
- We can calculate more low accuracy eigenvectors we need, and build up an eigenvalue preconditioner
- For degenerate eigenvalues (zero modes), we need to expand the projector $(1 - \mathbf{u}\mathbf{u}^\dagger)$ over our current estimate of the zero-mode subspace.
- The non-zero eigenvectors come in pairs, so by calculating ψ_i, then $\psi_{i+1} \sim (1 - \psi_i \psi_i^\dagger)\gamma_5 \psi_i$
- We need 3-4 inversions per eigenvector if we start from an accuracy $\sim 10^{-3}$.
- Jacobi-Davidson works well if we have a small number of eigenvectors calculated to a reasonable precision
- We require a reasonable initial guess to the eigenvectors, both for the deflation and to have a good starting (λ, u)

Zolotarev eigenvectors

- Basic idea: create a vector $b = \sum_i \psi_i$, where ψ_i are our best guesses of the eigenvectors
- Apply a step function $R = \frac{1}{2}(1 - \text{sign}(A - \lambda_0)\text{sign}(A + \lambda_0))$ to project b into the desired eigenvector subspace
- λ_0 lies between the largest eigenvalue we want to calculate to a high accuracy and the largest low accuracy eigenvector we possess
- Then use a Lanczos procedure to extract the wanted eigenvectors from $b' = Rb$
- In principle, we can use one set of inversions (on the same input vector) to calculate as many eigenvalues as we need.
- In practice, not as simple as this.

- We can see that ($b' = \sum_i \psi_i + \delta$)

$$\begin{pmatrix} F_1(A)b' & F_2(A)b' & F_3(A)b' & F_4(A)b' \end{pmatrix} - \begin{pmatrix} F_1(A)\delta & F_2(A)\delta & F_3(A)\delta & F_4(A)\delta \end{pmatrix} =$$

$$\begin{pmatrix} \psi_1 & \psi_2 & \psi_3 & \psi_4 \end{pmatrix} \begin{pmatrix} F_1(\lambda_1) & F_2(\lambda_1) & F_3(\lambda_1) & F_4(\lambda_1) \\ F_1(\lambda_2) & F_2(\lambda_2) & F_3(\lambda_2) & F_4(\lambda_2) \\ F_1(\lambda_3) & F_2(\lambda_3) & F_3(\lambda_3) & F_3(\lambda_3) \\ F_1(\lambda_4) & F_2(\lambda_4) & F_3(\lambda_4) & F_4(\lambda_4) \end{pmatrix}$$

$$(B - \Delta) = \Psi\Lambda$$

- F_i are arbitrary polynomial functions
- These need to be tuned so we use the smallest number of calls to the overlap operator to achieve $\|\Delta\Lambda^{-1}\| < \epsilon$
- We know (approximately) what the leading contributions to Δ , and we know what Λ is

- Choose $F_i = c_{in}(A/\lambda_*)^n$, for $n = 1, 2, \ldots N$
- Find the coefficients c_{in}, λ_* and N which minimise $8(\|\Delta\Lambda^{-1}\|)^3 + N$
- Then use those functions to efficiently extract the eigenvectors from b'
- To avoid degeneracies, we used the operator

$$A = \frac{1}{2}(1 \pm \gamma_5) \pm \frac{1}{2}(1 \pm \gamma_5)\textsf{sign}(K)\frac{1}{2}(1 \pm \gamma_5)$$

- This gives one of each eigenvector pair; we can easily reconstruct the second member of the pair

- There are difficulties (not yet fully resolved)
- Rounding errors limit the accuracy of the eigenvectors we can achieve
- Rounding errors also limit the number of F vectors we can usefully use
- The matrix sign function sign$(A \pm \lambda_0)$ can be approximated to a low but good enough accuracy by a Zolotarev rational approximation.
- We can use a multishift solver for the largest shifts, and switch to a deflated preconditioned eigCG inversion for the smaller shifts
- We require fewer inversions than the number of eigenvalues calculated each time
- In principle, this should beat Jacobi-Davidson to get the eigenvectors out to a moderate accuracy
- Finally, we can use Jacobi-Davidson to quickly polish the eigenvectors to a high accuracy if necessary.

- Number of Wilson calls for first n eigenvectors to converge to 10^{-9} precision
- N additional low accuracy eigenvectors.

n	N	Arnoldi	Jacobi-Davidson	Zolotarev
20	30	-	4.0×10^7	4.6×10^7

- These results are not final; both Jacobi-Davidson and Zolotarev can be improved
- We started in each case from eigenvectors with residuals between 10^{-6} and 10^{-2}.

Conclusions

- We have calculated bounds for the accuracy of the overlap operator required for a Arnoldi/SUMR eigenvalue routine to converge
- We need to use a high accuracy matrix sign for the entire Arnoldi calculation
- The convergence of the Arnoldi routine slows down considerably after a certain accuracy is reached
- The Jacobi-Davidson and Zolotarev routines can calculate eigenvectors to a high accuracy reasonably quickly using a low accuracy Dirac operator
- The Jacobi-Davidson routine currently wins on our test lattices
- We still have further optimisations to make, especially for the Zolotarev routine
- The Zolotarev routine seems to be particular sensitive to floating point errors, a problem we need to resolve